OPEN ACCESS

University of BRISTOL

Harris, R. (2016). Measuring segregation as a spatial optimisation problem, revisited: a case study of London, 1991–2011. *International Journal of Geographical Information Science*, *30*(3), 474-493. https://doi.org/10.1080/13658816.2015.1032973

Peer reviewed version

Link to published version (if available):
10.1080/13658816.2015.1032973

Link to publication record in Explore Bristol Research
PDF-document

This is the author accepted manuscript (AAM). The final published version (version of record) is available online via Taylor & Francis at http://www.tandfonline.com/doi/full/10.1080/13658816.2015.1032973. Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research
General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/

Measuring segregation as a spatial optimisation problem, revisited: a case study of London, 1991 – 2011

Abstract

There has been long and wide-ranging debate in the social science literature about how best to conceptualise and to measure segregation (Duncan & Duncan 1955a; Massey & Denton 1988; Morrill 1991; Allen & Vignoles 2007; Johnston & Jones 2010). A popular measure is the dissimilarity index, usually attributed to Duncan and Duncan (op. cit.) who were wise to its geographical limitations – that it, like most indices, is invariant to the precise spatial patterning of the segregation measured. Whilst one response to this shortcoming has been to develop a spatial adjustment (Morrill 1991), a number of papers from the 1980s and 90s took the approach of treating the measurement as a (spatial) optimisation problem (Jakubs 1981; Morgan 1983; Waldorf 1993). This paper revisits that optimisation literature, arguing that what was computationally prohibitive in the past is now possible in the open source software, R, and emblematic of the sorts of problem that might be more routinely solved in a cyberinfrastructure tailored to geographical analysis. Applying this method to UK Census data for London, and comparing the optimisation measure with the standard and adjusted dissimilarity indices, claims of ethnic desegregation are considered.

Key words: ethnicity, dissimilarity index, London, optimisation, segregation

Measuring segregation as a spatial optimisation problem, revisited: a case study of London, 1991 – 2011

1. Introduction

Segregation is the separation of one or more groups of people that have, or are given, characteristics that they or others imbue with particular meaning (for instance, race, religion, gender, wealth, age, social class) (Harris 2014b). How to measure segregation has been of enduring interest in sociology, and in urban and social geography. Morrill (1991, p.25) writes that "segregation is one of the most fundamental of human processes, and therefore its consistent and appropriate measurement is also important." Whilst educational research has looked at differences between schools – whether they are 'growing apart' and whether there is greater segregation between schools than between neighbourhoods (Gorard et al. 2003; Johnston et al. 2006; Johnston et al. 2007) – most studies of segregation rely on residential census data. Consequently, their focus has been on the degree to which different ethnic or social groups live in different places to one another.

The word segregation is associated with prejudice, discrimination and racism, especially in an ethno-cultural context. Without wishing to dismiss this meaning, nor to deny that such segregation has both historical and present-day occurrence (Nightingale 2012), such a strongly negative meaning is not intended here, nor in many of the other papers in which it is employed. The term is instead used as a synonym for the spatial separation of groups, without any a priori presumption about whether that separation is involuntary, caused by differential access to resources, social capital and to the housing and labour markets, whether it is in some way chosen by people (due to kinship networks, for example), and without judgement about whether the separations, where found, are necessarily undesirable (Merry 2011; Merry 2013).

There is long and wide-ranging debate in the social science literature about how segregation should be measured. A popular measure is the dissimilarity index, usually attributed to Duncan & Duncan (1955a). A limitation of the dissimilarity index, as with most indices, is that different geographical patterns of segregation generate the same index value. Whilst one response to this shortcoming has been a spatial adjustment to the index (Morrill 1991), a number of papers from the 1980s and 90s treat the measurement of segregation as a (spatial) optimisation problem (Jakubs 1981; Morgan 1983; Waldorf 1993), determining how far people would need to travel to nullify the patterns of segregation. This paper revisits that optimisation literature, arguing that what was computationally prohibitive in the past can now be undertaken in the open source software, R. It is anticipated that such software will be central to the emerging cyberinfrastucture which can be employed to undertake computationally demanding geographical analysis, of which spatial optimisation is an example.

The method, as well as the standard dissimilarity index and its spatial adjustment, is applied to an analysis of London over three decennial censuses,

1991, 2001 and 2011. Of these, 1991 was the first year the UK Census had a question about ethnicity; 2011 provides the most recent data. From the data, claims of ethnic desegregation are considered. London is of particular interest both because of its distinctiveness as a global city within the UK (Sassen 2001) – it has been described as both a different country and a de facto city state (O'Brien 2012; Jenkins 2013) – and because the changing demography of London has provoked debate both in the media and in academic writing (The Economist 2012; Goodhart 2014). Of note has been the numeric decline in those who classified themselves as White British from the categories of ethnicity available in the 2001 and 2011 Censuses – a decrease from 4.3 million to 3.7 million. This has occurred during a period when London's total population has increased from 7.2 million to 8.2 million. Consequently, whereas the White British comprised 59.8 per cent of the total in 2001, by 2011 they formed 44.9 per cent. Whether this is adjudged to be evidence of 'white flight' (Goodhart 2013; Hellen 2013), 'majority retreat' (Demos 2013), population mobility and the lure of the countryside (Easton 2013), greater population mixing (Catney 2013), the spatial de-concentration of minority groups (Johnston et al. 2014) or a demographic transition caused by higher birth rates amongst 'minority' ethnic groups as well as by a more aged White British population (Simpson et al. 2008; Greater London Authority 2013), the net result is to broaden London's historical identity as a multicultural city (Earle 1994) – albeit one where 63 per cent of Londoners were born in Britain, the population of London is 60 per cent white (including White British), and three-quarters of Londoners are British citizens (Katwala 2013).

The paper proceeds with further consideration given to how segregation is measured, focusing on the widely used dissimilarity index and its less widely used spatially adjusted counterpart. Section 3 introduces a literature on treating segregation as an optimisation problem, considering both the meaning of a measurement derived in this way and some practicalities in obtaining a solution. Section 4 provides the context for the analysis, reviewing the changing residential geography of ethnicity in London. Section 5 applies the dissimilarity index, the spatially adjusted index, and the optimisation measure to the same data and discusses the results. The conclusion follows with discussion about how the work is emblematic of the sorts of problem that might be more routinely solved in a cyberinfrastructure tailored to geographical analysis.

2. Measuring segregation

Evidence of segregation usually is treated as a form of social malaise, as the consequence of an unequal or divided society, and as having the potential to create distrust, a lack of mutual empathy and also misunderstanding between different social or ethno-cultural groups. Given its social and policy relevance, and its ability to generate emotionally charged debate, how best to measure segregation has been questioned and reviewed regularly within the academic literature. Morrill (1991, p.25) writes, "segregation is such a basic variable in social science that it deserves our best thinking and periodic rethinking."

A multitude of indices have been proposed to measure segregation, as have alternative methods including concentration profiles (see Section 3, below) and classifications of neighbourhood type (Poulsen et al. 2001; Johnston et al. 2002). A number of commentaries offer extensive reviews of these approaches, as well as their strengths and weaknesses (see, inter alia, Duncan & Duncan 1955a; Massey & Denton 1988; Morrill 1991; Gorard et al. 2003; Allen & Vignoles 2007; Peach 2009; Johnston & Jones 2010; Watts 2013; Harris et al. 2013). Here it is sufficient to note that most indices fall into one of two groups: those that have a probabilistic interpretation (e.g. the probability that a person selected at random from any neighbourhood will be of the same ethnicity as another person selected at random from the same neighbourhood: the isolation index, Shevky & Williams 1949; Bell 1954); and those that compare the distributions of two population groups across a study region, measuring the differences and constructing an average. Amongst the latter group and amongst all indices in general, the dissimilarity index is the most widely used. Gorard et al. (2003), talk of a 'Pax Duncana' crowning the dissimilarity index as the premier of all measures, although Duncan & Duncan (1955b) were actually more circumspect in their choice. The index is calculated as

$$D = 0.5 \sum_i \left| \frac{x_i}{N_X} - \frac{y_i}{N_Y} \right| \qquad [1]$$

where x_i is the number of people belonging to minority group X in neighbourhood i, y_i is the number belonging to a comparator group Y, N_X is the total count of group X across all neighbourhoods in the study region, and N_Y is the same for group Y. The index can be interpreted as the proportion of group X that would have to change location for the share of the group in each neighbourhood to be the same as the share of group Y. This interpretation assumes the second group does not move. The dissimilarity index ranges from zero (when the share of population group X is equal to the share of group Y in every location) to one (when X and Y never occupy the same places).

Sometimes the dissimilarity index is distinguished from a second segregation index calculated as

$$D = 0.5 \sum_i \left| \frac{x_i}{N_X} - \frac{(n_i - x_i)}{(N_T - N_X)} \right| \qquad [2]$$

where n_i is the total population in neighbourhood i, and N_T is the total population across all neighbourhoods. Equation 2 has the same form as Equation 1. The difference is in how the comparator group is defined. Equation 1 is used to compare the distribution of one population group (X) with one other (Y). Equation 2 is used to compare the one population group (X) with all others combined. However, Equation 2 simplifies to Equation 1 if we define $y_i = n_i - x_i \ \forall i$ and $N_Y = N_T - N_X$ (i.e. if Y represents all the population groups except X).

Another variant of the dissimilarity index is the index favoured by Gorard in his studies of social segregation within schools (Gorard et al. 2003) This compares

the distribution of group, *X*, with the distribution of the entire population including *X*:

$$G = 0.5 \sum_i \left| \frac{x_i}{N_X} - \frac{n_i}{N_T} \right| \qquad [3]$$

This index is directly correlated with the dissimilarity index (Allen & Vignoles 2007; Gorard 2009; Watts 2013). However, it has been criticised for what is described as double counting (since the amount x_i also contributes to n_i: Johnston & Jones 2010; Gorard 2011; Johnston & Jones 2011). This index will not be considered further, although *G* could be substituted for *D* in all that follows.

Other commentators have made the distinction between spatial and non-spatial indices of segregation (including White 1983; Morrill 1991; Harris 2011; Hong & O'Sullivan 2014). Most indices are judged non-spatial due to their assumption "that segregation can be measured without regard to the spatial pattern of [for example] white and nonwhite residence in a city" (Duncan & Duncan 1955a, p.215). Figure 1 illustrates the issue. To obtain 'maximum segregation' it does not matter how many cells are shaded black nor how many white, provided there is at least one of each. Equally, it does not matter whether cells of the same colour cluster together or not. In each case, the dissimilarity index reaches its maximum – correctly so in the sense that the white and black populations do not share a grid cell and are therefore fully separated. However, the spatial patterning clearly is not the same in each case. Intuitively, Scenario IX has the greatest segregation in that the two groups occupy the same amount of space but are arranged in a way that their distance apart is, on average, maximised.

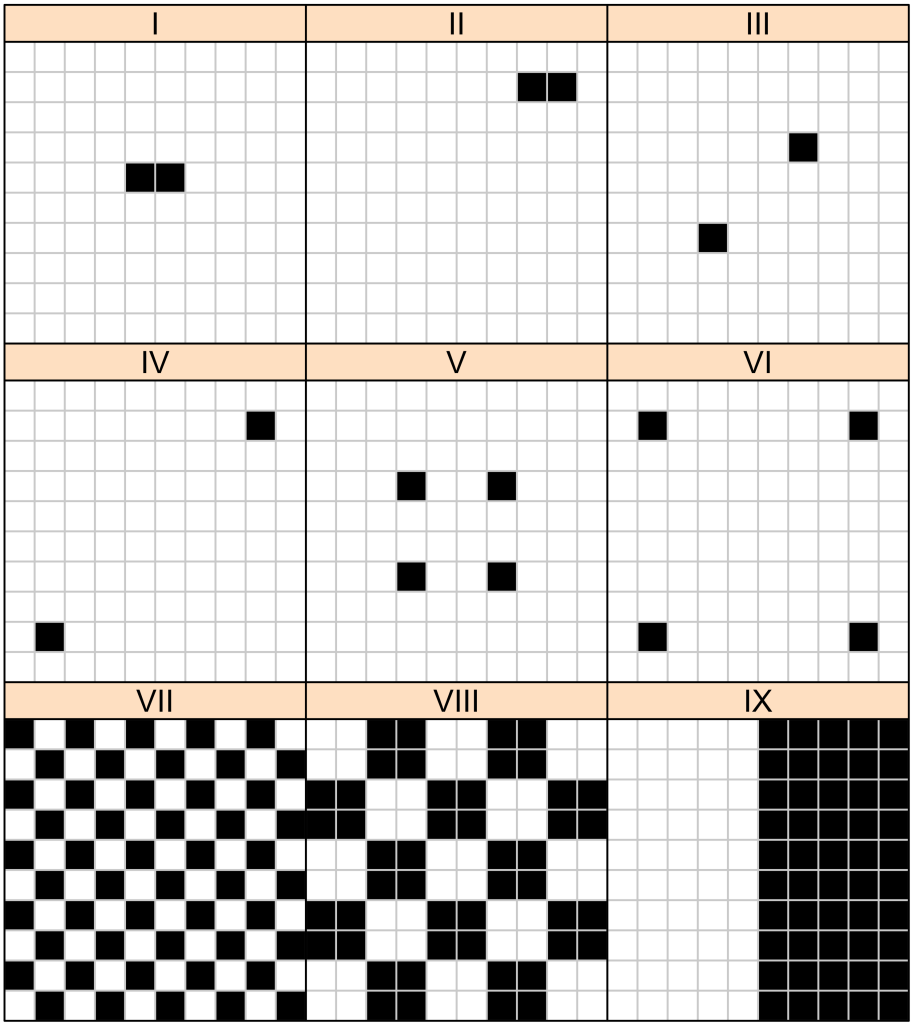

Figure 1. Hypothetical patterns of two populations to give, in each case, a maximum dissimilarity index value of one.

In response to this geographical shortcoming, spatial measures of segregation have been proposed (Wong 2003; Dawkins 2004; Reardon & O'Sullivan 2004). Of most immediate relevance is a modification to the dissimilarity index proposed by Morrill (1991):

$$D(adj) = D - \frac{\sum_i \sum_j w_{ij} \left| p_{(X)i} - p_{(X)j} \right|}{\sum_i \sum_j w_{ij}} \qquad [4]$$

given

$$p_{(X)i} = x_i / (x_i + y_i) \qquad [5]$$
$$p_{(X)j} = x_j / \left(x_j + y_j\right)$$

where w_{ij} are row-column entries in a spatial weights matrix, W. There are many ways the weights matrix could be defined, including by the inverse of the

distance between any two neighbourhoods (grid cells in Figure 1; census reporting zones for the majority of this paper) or by the length of a shared boundary (Wong 1993). More simply, contiguity can be employed: places that share a boundary are assigned a weight of 1; those that do not are assigned 0. Given this, recognising that $p_{(X)i}$ is the proportion of the relevant population that belongs to group X in location *i*, and also recognising that where $w_{ij} = 1$, $w_{ij}\left|p_{(X)i} - p_{(X)j}\right|$ is the difference in the proportions of *X* for two adjacent locations, then what the adjustment does is subtract from *D* the mean difference in proportions for all contiguous neighbourhoods. The logic of the adjustment is described by Morrill: "if a very high proportion of the common boundaries with other tracts show a similarly high or low percent minority [as in pattern IX in Figure 1] then *D* is meaningful as it is, because there are limited opportunities to interact across space; but if a high proportion of the common boundaries show a big minority-majority difference [as in pattern VII], a high degree of opportunity to interact across space is present" (p. 34).

Table 1 reports the *D(adj)* scores for each of the scenarios shown in Figure 1, using a contiguity matrix and considering both the Rook's case, when diagonals are not considered to be neighbours, and The Queen's case, when they are. The bottom line of the table presently should be ignored. An error of interpretation arises for Scenario VII in the Rook's case where a zero value is obtained from the spatial adjustment, implying no segregation. This happens because the difference in the proportions is always one between adjacent cells and because the sum of those differences is equal to the sum of the weights. Consequently, the numerator and denominator of the adjustment term are equal (in Equation 4), yielding $D(adj) = D - 1 = 1 - 1 = 0$. However, the circumstance in which this arises is remarkably contrived, requiring a checkerboard arrangement on a regular grid. It is not a realistic real-world scenario.

In other respects, the adjustment acts according to intuition. For example, scenarios I and II have greater clustering of the black group than do III and IV and their index scores now reflect it. The degree of clustering is the same in I and II and their adjusted scores are equal. The actual location of the cluster is not the same in both I and II, however the score is invariant to that change (as it is also to the locational differences of III versus IV or V versus VI). In comparison to IX, scenarios VII and VIII have the black group distributed widely across the study region and this leads to a much lower adjusted score.

Scenario	I	II	III	IV	V	VI	VII	VIII	IX
D	1	1	1	1	1	1	1	1	1
D(adj) rook's case	0.967	0.967	0.956	0.956	0.911	0.911	0	0.556	0.944
D(adj) Queen's case	0.959	0.959	0.953	0.953	0.906	0.906	0.474	0.532	0.918
D(op)	3.599	5.504	3.079	3.860	2.310	2.310	1.000	1.118	5.000

Table 1. Measures of dissimilarity obtained under the different spatial configurations shown in Figure 1.

Hong & O'Sullivan (2014) cite the adjusted dissimilarity index as an example of what they call a zone-based approach. This they contrast with a surface-based approach that treats the measurement at location *i* as part of a continuous surface of values changing all across the study region and not just at the boundaries of neighbourhood zones. Interpolation techniques are employed to model these surfaces with parallels between this approach and geographically weighted statistics (Fotheringham et al. 2002): both create localised statistics based on comparing a location with other locations around it, and both view space as a continuous field rather than a tessellation of discrete zonal objects.

As an example of a surface-based approach and within a wider exposition of spatial measures of segregation, Reardon and O'Sullivan (2004) outline what they describe as a spatial dissimilarity index, for which readers are referred to Equation 12 of their paper. That index is interpreted "as a measure of how different the composition of individuals' local environments are, on average, from the composition of the population as a whole" (p.141). Implicit to a surface-based approach is the idea that levels of segregation vary smoothly (therefore continuously) across the study region. Other have taken the opposite view, treating segregation as a spatial discontinuity/disparity between neighbouring locations (Chakravorty 1996; Mitchell & Lee 2014; Harris 2015). A further approach exploring the spatial properties of segregation is by measurement at multiple scales (Lee et al. 2008; Lloyd 2012).

3. The measurement of segregation as a spatial optimisation problem

In addition to the modifications to the dissimilarity index discussed above, there has been periodic interest in treating the measurement of segregation as an optimisation problem. The task is then to determine the average distance members of a population group would need to travel to achieve 'no segregation' (Jakubs 1981; Morgan 1982, 1983; Waldorf 1993). The idea works by decomposing the index of dissimilarity (or some related index) into its component parts,

$$D \propto |d_1| + |d_2| + |d_3| + \cdots + |d_n| \qquad [6]$$

where

$$d_1 = \frac{x_1}{N_X} - \frac{y_1}{N_Y}, \quad d_2 = \frac{x_2}{N_X} - \frac{y_2}{N_Y}, \quad \text{etc.} \qquad [7]$$

The numeric subscripts (1, 2, etc.) each represent a location, and at each location there will be a surplus (Sp) or deficit (Df) in the share of population group *X* relative to *Y*:

$$Sp = x_i - \frac{y_1 N_X}{N_Y} \quad \forall \ \frac{x_i}{N_X} > \frac{y_i}{N_Y} \qquad [9]$$

$$Df = \frac{y_1 N_X}{N_Y} - x_i \quad \forall \ \frac{x_i}{N_X} < \frac{y_i}{N_Y} \qquad [10]$$

Given a cost matrix where the penalty is the Euclidean distance from any one location to another, the least cost (minimum distance) solution is determined, that, by moving the minority population, *X*, from one location to another would give a dissimilarity score of zero. This least cost solution, D_{MIN}, is determined using linear programming and by solving with direct analogy to the transportation problem (Monge 1781).

Following Jakubs (1981) the problem formally is specified as follows. The origins, O_i, are the locations with a surplus of the group:

$$O_i = x_i - \frac{y_1 N_X}{N_Y} \quad \forall \ \frac{x_i}{N_X} > \frac{y_i}{N_Y} \qquad [11]$$

and the destinations, D_i, are those with a deficit:

$$D_i = \frac{y_1 N_X}{N_Y} - x_i \quad \forall \ \frac{x_i}{N_X} < \frac{y_i}{N_Y} \qquad [12]$$

The task is to minimise,

$$z = \sum_{ij} X_{ij} c_{ij} \qquad [13]$$

where X_{ij} is the number of group *X* that would have to move from location *i* to *j* to produce a dissimilarity index score of zero, and c_{ij} indicates the cost of doing so (which is the distance), subject to:

$$\sum_i X_{ij} = D_j \text{ for all } j \text{ zones} \qquad [14]$$
$$\sum_j X_{ij} = O_i \text{ for all } j \text{ zones} \qquad [15]$$
$$X_{ij} \geq 0 \qquad [16]$$

Having determined D_{MIN}, a concern of earlier writers was how to convert it into an index value ranging between 0 and 1. Their solution was to compare the distance with the value obtained under some hypothetical scenario of maximum segregation: when, for example, the entire population group *X* is assigned to the most remote locations from *Y*. This has the drawback of being computationally demanding: not only must the minimum distance be determined twice (once for the actual population distribution, the second for the hypothetical distribution) some sort of population reallocation algorithm is required to construct the maximum segregation circumstance.

In any case, the least distance solution offers a readily interpretable value without conversion to an index score. Dividing the distance by N_X gives the

average distance a member of group *X* would have to travel to achieve 'no segregation'. A problem is that the average distance is affected by the geography of the study region. It is expected to be higher in regions where the distances between the locations are greater – for example, more so in rural areas than in cities, or in suburban areas than in areas of high density housing (because census tracts are generally smaller in areas where the population density is greater).

There are three responses to this problem. A potential option is to calculate the mean distance between the census tracts (the mean inter-centroid distance, for example) and divide D_{MIN}/N_X by it. That would act to 'standardise' the measure with consideration to how far the zones are apart. A second possibility would be to replace the Euclidean distances with nearest neighbour distances (distance to the first nearest neighbour is 1, to the second, 2, and so forth). That is not pursued here although initial testing showed the optimisation score remains affected by the spatial configuration of the study region: more compact shapes have more accessible neighbours whether distances are measured using a Euclidean or nearest neighbour metric. A third possibility is to concede that different study regions are not directly comparable, although any study region can be compared with itself over time. That is the approach used here where the focus is on a single study region, with a fixed boundary, and where the aim is not to compare different places on one occasion but one place for three occasions – London, in 1991, 2001 and 2011. It may be noted that even with conventional segregation indices, the comparison of different study regions (for example, local authorities, government regions or cities) is far from straightforward when those places come in different shapes and sizes, some bigger, some smaller, some with high population densities, and some with low, which means the places are not directly comparable especially when the indices are sensitive to those differences (Wong 1997).

A separate issue is what the distance-based index of dissimilarity is measuring. In their seminal paper on the dimensions of residential segregation, Massey and Denton (1988) categorise the optimisation value (in index form) under the dimension of clustering. Their other dimensions are evenness (which the standard dissimilarity index is said to measure), exposure, concentration and centralization.

The categorisation is correct only in part. To see why, look again at Table 1 where the bottom row provides the average distance measure, $D(op) = D_{\mathrm{MIN}}/N_X$, calculated for each of the spatial arrangements shown in Figure 1. For the calculations it is assumed that the white and black populations are spread evenly across all the grid cells where they have a presence (i.e. x_i/N_X is a constant for all cells where $x_i > 0$, the black shaded cells, and y_i/N_Y also is a constant where $y_i > 0$, the white shaded cells). Like the spatially adjusted measure, *D(op)* regards Scenario IX as having greater segregation than VII, and both more so that VI. That fits with intuition. However, unlike the spatially adjusted measure, *D(op)* is sensitive to where the clusters are located: the value for II is greater than for I because the clusters are further from the centre; IV yields a greater value than III for the same reason.

It might be suggested therefore that *D(op)* is a measure of (de-)centralisation, rising in value the less central the clustering is. However, the truth is subtler than that. Notice, for example, that Scenario I has a more centralized pattern than Scenario III but a higher *D(op)* score. Scenarios V and VI have the same score but the former is more centralized. What *D(op)* actually measures is accessibility – how easily (in distance terms) the areas with a surplus share of population group *X* can be reached from areas with a deficit. In other words, it is a geographical measure of the residential separation of group *X* from *Y*. If it is assumed that distance affects contact or interaction, then it is a measure of inter-group exposure.

In fact, Massey and Denton had little opportunity to test the distance-based measure empirically. They write, "After some initial experimentation […] we discovered that computing costs and machine requirements quickly became prohibitive for large urban areas" (ibid., p.296). However, it is not really the size of the study area that matters *per se* but the number of sub-divisions within it. Massey and Denton were using census tracts with an average population size of between 3000 and 6000 people for fifty of the largest Standard Metropolitan Areas (SMAs) in the United States, as well as some additional areas. Considering that the population count for New York alone was 7 million in 1980, and that includes neither the population of Newark nor New Jersey which are in the same SMA, then it is not surprising that Massey and Denton found the method to be computationally intractable at the time. Jakub's (1981) proof of concept was limited to testing on an 8 by 8 grid and an application to 183 census tracts in Marion County, Indiana. Morgan's (1983) implementation was tested on a 16 by 16 grid and applied to a study of 354 tracts in Philadelphia. To the best of the current author's knowledge, the method has never before been implemented for a larger number of areas.

Although the computational demands are now less prohibitive, they are not immaterial. To demonstrate this, a regular grid of *k* cells was generated. Values of *X* and *Y* were arranged in a checkerboard pattern with a zero count of *Y* in every second cell, and a zero count of *X* in between (giving a dissimilarity index value of one). Non-zero values of *X* and *Y* were drawn from a random uniform distribution with minimum zero and maximum $4\,N/k$ where *N* denotes either N_x or N_y. This gives an expected total of N_X = 500 000 for *X*, and N_Y = 8 000 000 for *Y*. The value of *k* was increased stepwise with values of 36, 144, 256, 484, 676, 1024, 1296, 1764 and 2116.

The times taken to solve the optimisation problem are shown in the second column of Table 2. It takes almost 4 million times longer to determine the answer for 2116 grid cells than for 36 although their number has increased only by a factor of 60. Furthermore, the times to determine a solution are a function of the spatial patterning of the two populations. In the current case there is the checkerboard pattern of strongly negative spatial autocorrelation. The times sharply increase if the pattern is changed to one of strong positive autocorrelation. To show this, each of the cells was assigned a value of *X* and *Y*, where the values were drawn from a random uniform distribution with minimum zero and a maximum of $2\,N/k$ (giving the same expected totals as

before) and ordered so that the cells closest to the bottom-left corner of the grid received the highest count of *X*, those furthest away the least, and vice versa for *Y*. The new times taken for the solutions are in the third column of Table 2. For the 2116 grid cells it has increased to 44 minutes. A standard laptop was used (a 2-year-old MacBook Pro).

Morrill (1991, p.30) describes the set-up required to measure segregation as a transportation problem as "rather tedious". That, at least, is no longer true. Here the calculations are undertaken in the open source statistical and computing software, R. The scripts are simple, calling on R's lpSolve library to determine the solution, and the maptools, sp and spdep libraries to handle the spatial data and to calculate the distances matrix. The lpSolve library is an interface to a free linear (integer) programming solver based on the revised simplex method and the Branch-and-bound method for the integers (Mitchell, 2002). Full documentation can be found at http://lpsolve.sourceforge.net/. No claim is made that this is necessarily the fastest way of solving the optimisation problem. The purpose of the current paper is to show only that the distance-based approach to measuring segregation is now a realisable concept for a large city such as London and can be implemented with minimum specialised knowledge, using 'off-the-shelf' libraries in R.

Number of grid cells, k	Strong negative spatial autocorrelation	Strong positive spatial autocorrelation
36	0.003 sec	0.003 sec
144	0.064 sec	0.122 sec
256	0.405 sec	0.861 sec
484	2.44 sec	8.87 sec
676	6.80 sec	40.7 sec
1024	27.1 sec	3.36 min
1296	57.6 sec	5.11 min
1764	1.92 min	28.4 min
2116	5.95 min	44.3 min

Table 2. Time taken to solve the optimisation problem in R given the number of grid cells, *k*, and under patterns of strongly negative and strongly positive spatial autocorrelation

4. Geographies of Ethnicity in London, England

Figure 2 sets the demographic context for the analyses that follow. It shows what percentage of the total London population is given by eight of the most populous ethnic group categories in each of the 1991, 2001 and 2011 Censuses of England and Wales. The White British group is a subcategory of all white ethnicities and was introduced in 2001, as were the mixed ethnicity categories (shown aggregated together in Figure 2). Each of the minority groups has increased its share of the London population, with the exception of Black Caribbeans who have remained at about 4 per cent of the total. The white population has decreased its share of the total, from 80 per cent in 1991, to 60 per cent in 2011.

A part of the decline may be attributed to the introduction of the mixed ethnicity categories. However, even if we subtract from the 2001 white group total the number who described themselves as of a mixed white and other ethnicity in 2011, then the decline in the white group share is still from 67 per cent in 1991.

Furthermore, it is not only in relative terms that the white population has declined. There were fewer people of a white ethnicity counted in 2001 than in 1991, and fewer again by 2011. Against an increase in London's total population of 7 per cent from 1991 to 2001, and a further 14 per cent from 2001 to 2011, the number of white population fell by four per cent from 1991 to 2001, and by the same amount from 2001 to 2011 (during which the White British population fell by 14 per cent). Each of the other groups shown in Figure 2 have grown in number over the same periods, although the rate of growth is faster from 1991 to 2001 than from 2001 to 2011 (the groups are starting at a lower base in 1991). Overall, the non-white poulation count has increased by 144 per cent from 1991 to 2011 in London. It was 1.35 million in 1991, 3.29 million in 2011.

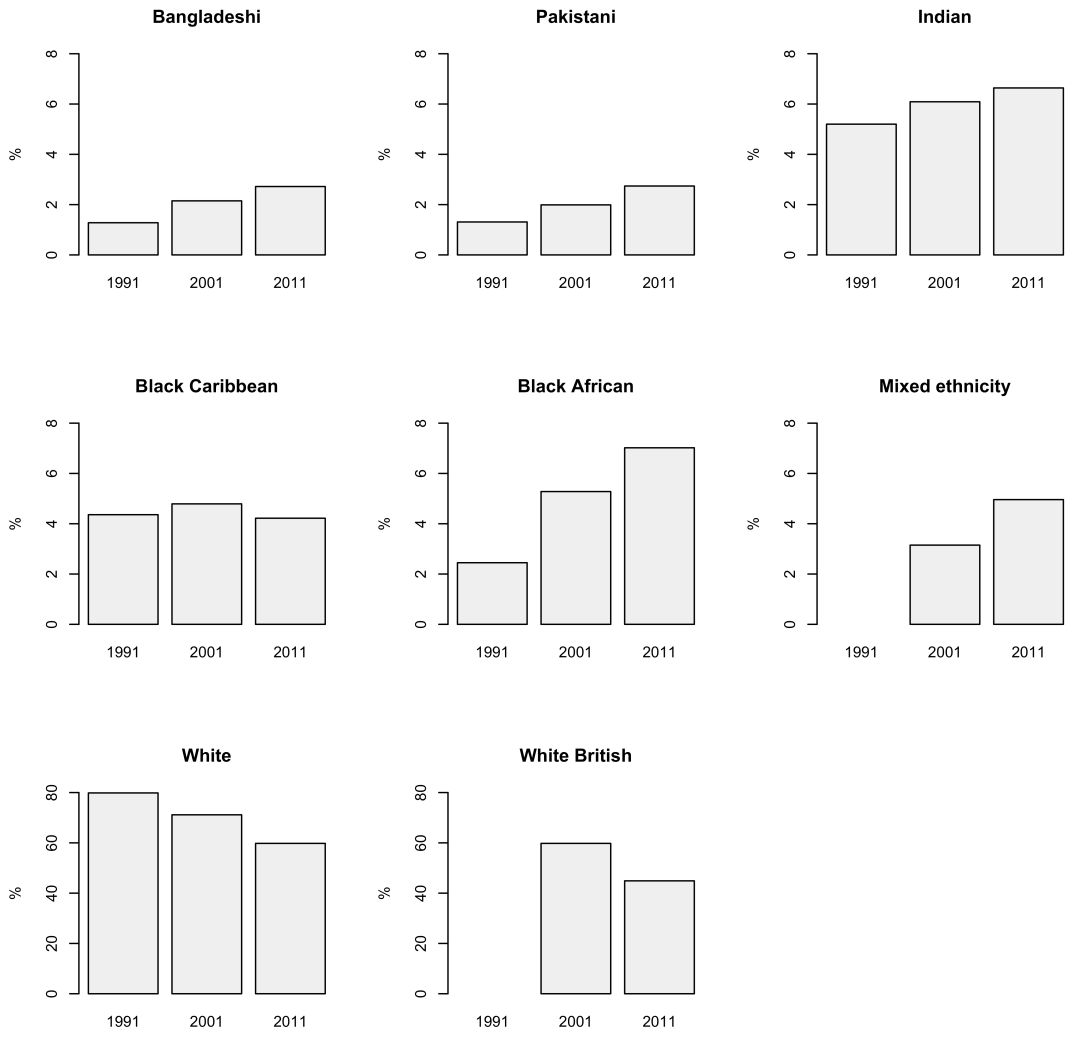

Figure 2. The percentage of the population belonging to various ethnic groups in London according to the 1991, 2001 and 2011 Census data.

Figure 3 maps the geography of non-white population for London from 1991 to 2011. Specifically, it maps the percentage of the residential population in London neighbourhoods that classified themselves of an ethnic group other than white in the 1991, 2001 and 2011 UK Censuses. The neighbourhoods are the 2011 Lower Level Super Output Areas (LLOAs), of which there are 4800 with a mean average population of 1690 persons (median 1654, interquartile range from 1530 to 1817, full range from 985 to 4933).

LLOAs are the second tier census geography in England and Wales. They are aggregates of the finer scale Output Areas (OAs) and are used here for two reasons. First, to provide a consistent geography. Between 1991 and 2001, the census geography for England and Wales was redesigned with what were then called Enumeration Districts (ED) completely redrawn as Output Areas, and Super Output Areas introduced for the first time. Although there has been less change between 2001 and 2011, some OAs have been split or merged according to local population change. Consequently, the greatest change to OAs occurs in the places of most demographic interest, where population change is greatest. To estimate the LLOA population counts for each census, the finest scale ED/OA data have been aggregated into the 2011 LLOA geography by a point-in-polygon assignment, where the point is the small area centroid.

The second reason for choosing LLOAs is computational. Table 2 showed how the time taken to solve the optimisation problem scales poorly with the number of areal units, k. Using LLOAs for London gives $k \cong 4800$. Choosing OAs would give $k \cong 25000$.

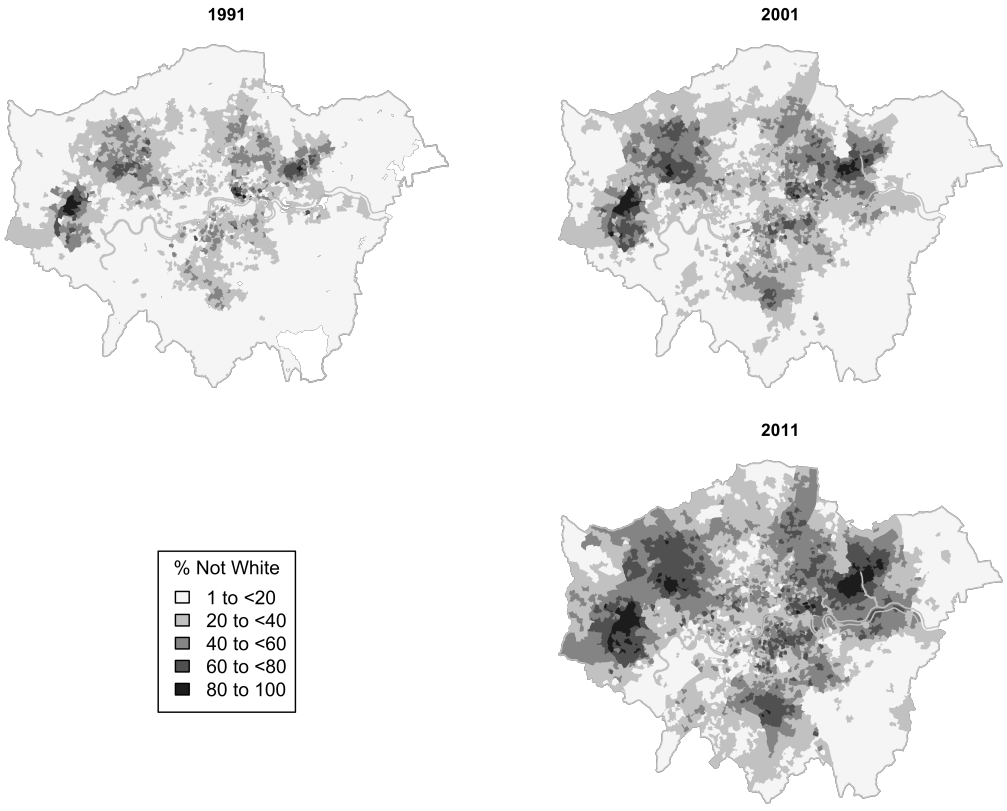

Figure 3. Showing the percentage of the population of a non-white ethnicity in London in 1991, 2001 and 2011.

Figure 3 suggests that the non-white groups are spreading out over larger areas of London. Certainly they form a rising percentage of the residential population in most London neighbourhoods over the twenty year period. Overall, 97 per cent of LLOAs had a greater percentage of non-white population by 2011 than in 1991. The mean increase is nineteen percentage points, with an average of 40 per cent of the local population being of a group other than white by 2011. The raised percentages could be driven by the decline of the White population not because the non-white groups are themselves becoming more spread out. However, direct evidence for the geographical deconcentration of the non-White groups is provided by Figure 4. It gives what are known as concentration profiles (Johnston et al. 2003), showing, for example, that if we rank LLOAs from highest to lowest in regard to the percentage of their population who are Bangadeshi, then 70 per cent of London's entire Bangadeshi population are located in less than 20 per cent of its neighbourhoods, and 95 per cent are located in just 40 per cent of all neighbourhoods in 2011. By contrast, the Black African group is more geographically spread: 70 per cent of their population is found in about 30 per cent of neighbourhoods. Note that the values increase downwards along the vertical axes. This allows the charts to be arranged so that the higher the curve, the greater the geographical concentration. That the white population is shown to be least concentrated (most spread-out) is hardly suprising, it is the largest group. Note that for most groups the level of geographical concentration has decreased from 1991 to 2001, and again from 2001 to 2011 (the curve is moving downwards). The exception is the White group who are becoming more concentrated in particular neighbourhoods.

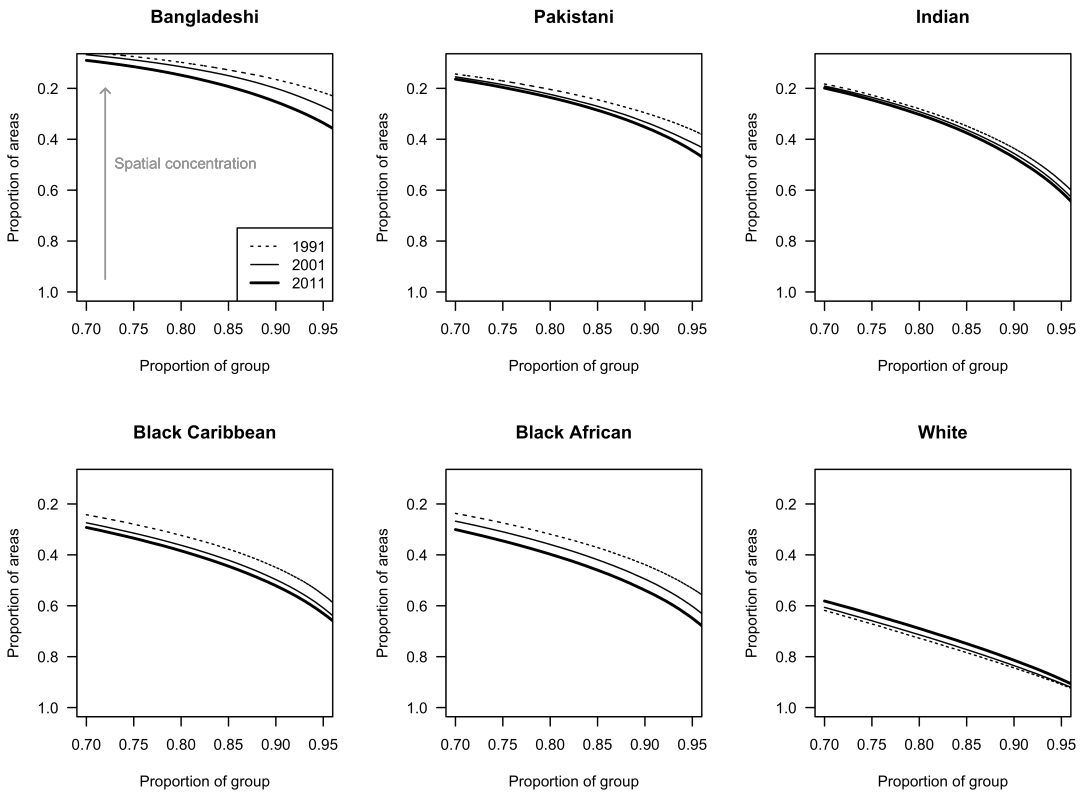

Figure 4. Concentration profiles for six ethnic group categories present in the 1991, 2001 and 2011 Census data for London.

5. Applying the *D*, *D(adj)* and *D(op)* measures to London

Table 3 provides the values of the dissimilarity index and its spatial variants for each of the six ethnic groups included in Figure 4, calculated for London using the 1991, 2001 and 2011 Census data at the LLOA scale. The values are calculated where *X* is the ethnic group named in the column headings and where *Y* is the remainder of the population (i.e. the version of *D* described by Equation 2, above). *D(adj)* is calculated using contiguity matrices and the Queen's case definition of neighbouring. *D(op)* requires the distances between LLOAs to be known. Their inter-centroid distances are used.

	Bangladeshi	Pakistani	Indian	Black African	Black Caribbean	White
1991						
D	0.674	0.538	0.510	0.449	0.466	0.395
D(adj)	0.663	0.529	0.485	0.435	0.446	0.329
D(op)	7.40	3.96	6.23	4.26	3.78	3.31
2001						
D	0.649 (🡻)	0.502 (🡻)	0.492 (🡻)	0.414 (🡻)	0.426 (🡻)	0.382 (🡻)
D(adj)	0.636 (🡻)	0.492 (🡻)	0.467 (🡻)	0.388 (🡻)	0.409 (🡻)	0.310 (🡻)
D(op)	7.68 (🡹)	3.92 (🡻)	6.32 (🡹)	3.64 (🡻)	3.36 (🡻)	3.05 (🡻)

2011						
D	0.617 (🡻)	0.505 (🡹)	0.486 (🡻)	0.378 (🡻)	0.385 (🡻)	0.356 (🡻)
D(adj)	0.603(🡻)	0.494 (🡹)	0.461 (🡻)	0.347 (🡻)	0.370 (🡻)	0.275 (🡻)
D(op)	7.71 (🡹)	4.17 (🡹)	6.25 (🡻)	3.28 (🡻)	3.06 (🡻)	2.65 (🡻)

Table 3. The dissimilarity measures for the six main ethnic groups in London in 1991, 2001 and 2011.

Generally the three dissimilarity measures are correlated strongly with each other. The Pearson correlations in Table 4 are obtained by pooling together the *D* scores for all groups and all years, and doing the same for the *D(adj)* and *D(op)* scores. The log of *D(op)* is taken because the distance measure is skewed (its upper bound is limited only by the size of the study region). The correlation between the conventional dissimilarity index and the spatially adjusted version is especially strong – almost perfect – suggesting that although *D* may not be a true spatial measure of segregation, it nevertheless performs well in the real world example of London.

	D	*D(adj)*	$log_e D(op)$
D	-	0.986	0.813
D(adj)	0.986	-	0.810
$log_e D(op)$	0.813	0.810	-

Table 4. The correlations between the dissimilarity measures

The correlation between the two spatial measures, *D(adj)* and *D(op)* is high but not as great suggesting there is greater differentiation in what they measure. Regressing $log_e D(op)$ against *D(adj)* and looking at the residuals permits the occasions to be determined when *D(adj)* most under-predicts, most over-predicts and when it best fits *D(op)*. Those occasions are, respectively, the Indian group in 2011, the Pakistani group in 1991, and the Bangladeshi group in 1991.

Figure 5 maps the residential geography of those groups in the years concerned. In the case of $log_e D(op)$ being under-predicted, the situation is similar to II in Figure 1, albeit with more than one cluster. Crucially, these clusters of Indians are towards the edge of the study region, raising the distance metric. In the example of over-predication, the situation is more akin to the hypothetical scenarios V and VI – the clusters of Pakistanis are distributed in relatively accessible locations (and the clusters are less pronounced). In the best-fit case, there is a relatively small cluster of Bangladeshis, located towards the centre of the study region, which is, by definition, of average accessibility.

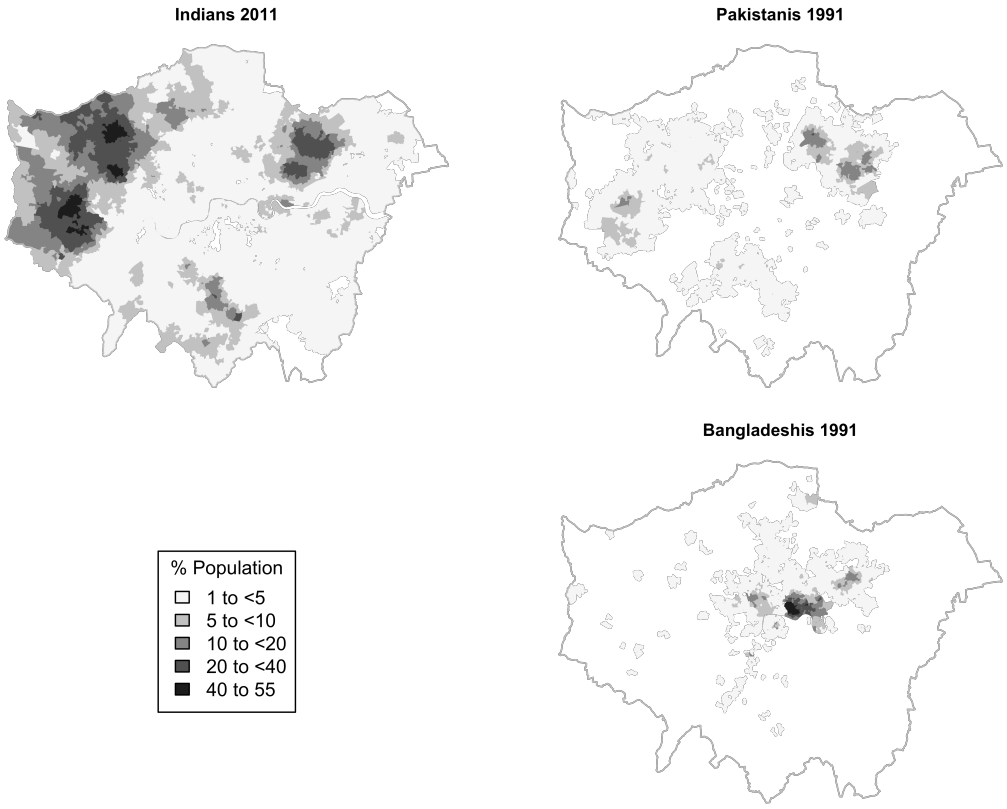

Figure 5. (top left) when *D(op)* is most under-predicted by *D(adj)*; (top right) when it most over over-predicted; (bottom right) when there is the best fit.

Recall that *D* is said to measure unevenness, *D(adj)* spatial clustering and *D(op)* is a measure of inter-group exposure (within the boundaries of the London region). On each of these dimensions, the overall trend is one of decreasing ethnic segregation within London, which is consistent with the findings of other studies (Catney 2013; Harris 2014a; Johnston et al. 2014). A slight exception to the trend is the Pakistani group, which appears to be a little more segregated in 2011 than in 2001 but still less so than in 1991.

On the optimisation measure, the Bangladeshi group is unusual in being the only group for which the distance has increased in value at each successive census. Yet the *D* and *D(adj)* values have decreased for this group over the same period, as has the concentration profile. A logical explanation is that the group has expanded into neighbourhoods that are towards the edge of London. Figure 6 confirms that to be correct. It shows the population-weighted mean centre of the Bangladeshi group in 1991, 2001 and 2011. The shift is almost 1km eastwards, on average over the period, with the Bangladeshis forming an increasing percentage of the residential population of northeast London.

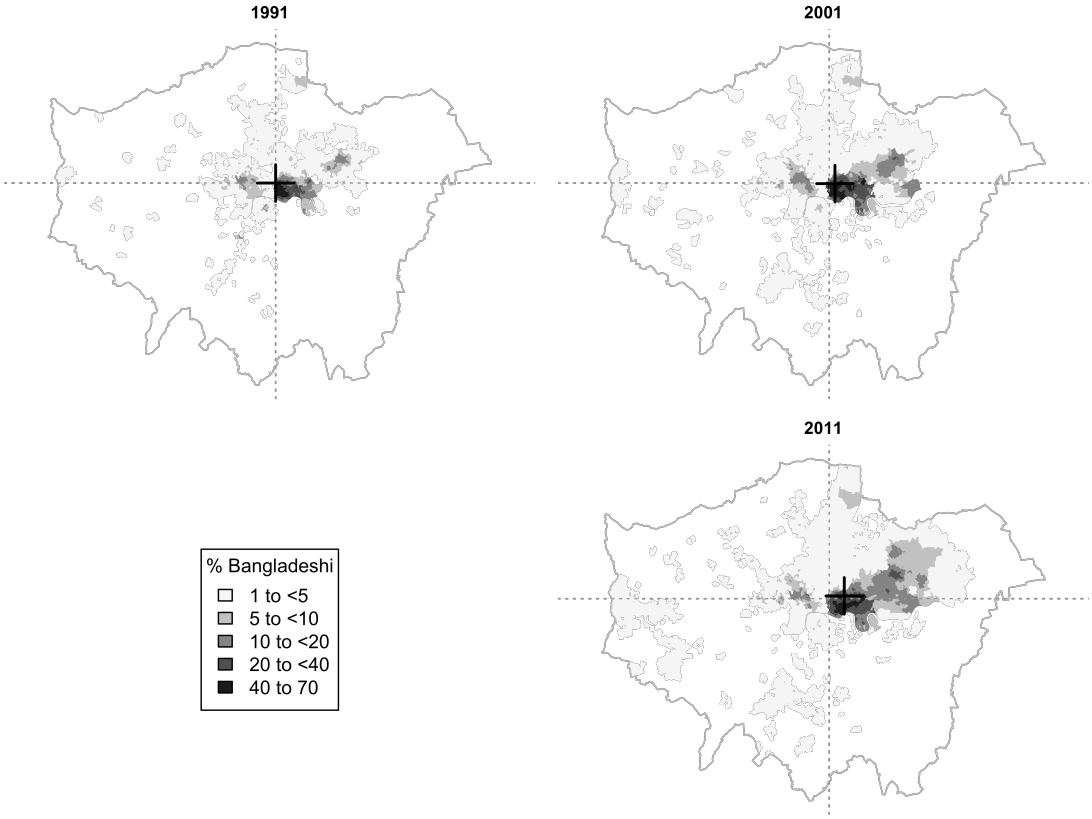

Figure 6. Showing the residential geography of the Bangladeshi group in 1991, 2001 and 2011. The mean centre in each year is shown as a cross. The movement of the group is eastwards.

In the case of the white group, we cannot ignore their decreased number in London or that those who remain are more concentrated in particular neighbourhoods than they were before. We may suspect that the decreasing segregation scores are caused by other ethnic groups moving into areas where the number of the population classified as white, and especially White British, is declining: a process that creates more mixed neighbourhoods in London. Notwithstanding its changing ethnic composition, it remains the case that London's residentially population is less ethically segregated than it was.

6. Conclusions

This paper has revisited a literature that treats the measurement of segregation as an optimisation problem. It has applied the method to look at residential segregation in London, by ethnicity, over three successive censuses. It has been implemented in the open source software, R.

There are two sets of conclusions, one socio-demographic, the other methodological. With regard to the former, the overwhelming evidence is that ethnic segregation has declined in London over the twenty-year period to 2011. Three factors have driven this: the growth in number of the 'minority' population

groups, their spreading-out across the city but also the decline and spatial retrenchment of the white population, notably the White British. The net result is more mixed neighbourhoods albeit with a lower prevalence of the majority (White British) group within the city.

With regard the latter, the paper has argued that the computational difficulties that all but prohibited the adoption of the optimisation approach in the past have largely passed, although the time taken to determine a solution is not trivial, especially given a large number of data zones and strong patterns of positive spatial autocorrelation. What the approach produces is a measure of accessibility between ethnic groups, given in terms of a distance metric. This metric is sensitive not only to the degree of spatial clustering of population groups but also to where in the study region the clustering is found. This differentiates the optimisation measure from Morrill's spatially adjusted measure (Morrill 1991). As a consequence of the sensitivity, it would be unwise to use the optimisation approach as a standalone measure of segregation as an interpretive error can arise. It is important to be aware that a rise in its value need not indicate an increase in segregation but could be due to the spreading-out of an ethnic group into the more peripheral parts of a city, into neighbourhoods that actually become more mixed.

However, it is equally unwise to focus solely on the conventional dissimilarity index or its spatially adjusted counterpart when considering changing patterns of segregation. Segregation is a multi-faceted, spatial process that should not be reduced to a single index or classification (Massey & Denton 1988). Handled sensitively and used together in conjunction with other contextual information, the three dissimilarity measures outlined in this paper can provide more complete information on changing patterns of (ethnic) segregation within a city – specifically, information about unevenness, clustering and between group accessibility / exposure.

More generally, what the paper illustrates is the increased tractability of computationally demanding geographical methods of analysis using open source software. The same software can be used to undertake local methods of analysis, including geographically weighted statistics that also are computationally demanding for large data sets; and it is not restricted to a standalone laptop but can be implemented to run using a server. It is here that the development of software such as R within a cyberinfrastruture can be anticipated, providing a point-of-entry (e.g. graphical interface) for users who are interested in the analysis but who wish to avoid the scripting, and further reducing the time required to complete the analysis given a suitable server and computing infrastructure. Harris et al. (2010) demonstrate how geographically weighted regression could be implemented in R using Grid computing. More recent developments such as R Studio Server and Shiny (www.rstudio.com/products/shiny/), as well as the libraries for handling geographical information documented by Bivand et al. (2013) and Brunsdon and Comber (2015), and the library containing a suite of segregation measures authored by Hong and O'Sullivan (2014), offer the prospect of building a geographically tailored cyberinfrastruture for handling geographical information

and using it to address issues of an applied, geographical nature, of which the enduring interest in how best to measure and model processes of segregation, is emblematic.

References

Allen, R. & Vignoles, A., 2007. What Should an Index of School Segregation Measure? *Oxford Review of Education*, 33(5), pp.643–668.

Bell, W., 1954. A Probability Model for the Measurement of Ecological Segregation. *Social Forces*, 32(4), pp.357–364.

Bivand, R.S., Pebesma, E. & Gómez-Rubio, V., 2013. *Applied Spatial Data Analysis with* R (2nd edn.). Berlin: Springer.

Brunsdon, C. & Comber, L., 2015. *An Introduction to R for Spatial Analysis and Mapping*. London: Sage.

Brunsdon, C., Fotheringham, A.S. & Charlton, M., 2002. Geographically weighted summary statistics – a framework for localised exploratory data analysis. *Computers, Environment and Urban Systems*, 26, pp.501–524.

Catney, G., 2013. *Has Neighbourhood Ethnic Segregation Decreased? The Dynamics of Diversity: evidence from the 2011 Census*, University of Manchester: Centre on Dynamics of Ethnicity.

Chakravorty, S., 1996. A Measurement of Spatial Disparity: The Case of Income Inequality. *Urban Studies*, 33(9), pp.1671–1686.

Dawkins, C., 2004. Measuring the spatial pattern of residential segregation. *Urban Studies*, 41(4), pp.833–851.

Demos, 2013. Almost half of ethnic minority population now live in majority non-white areas. Available at: http://www.demos.co.uk/press_releases/almosthalfofethnicminoritypopulationnowliveinmajoritynonwhiteareas [Accessed July 22, 2014].

Duncan, O.D. & Duncan, B., 1955a. A Methodological Analysis of Segregation Indexes. *American Sociological Review*, 20(2), pp.210–217.

Duncan, O.D. & Duncan, B., 1955b. Occupational stratification and residential distribution. *Journal of Sociology*, 60(5), pp.493–503.

Earle, P., 1994. *A City Full of People: Men and Women of London, 1650-1750*, York: Methuen Publishing Ltd.

Easton, M., 2013. Why have the white British left London? *BBC News online*. Available at: http://www.bbc.co.uk/news/uk-21511904 [Accessed July 22, 2014].

Economist, The, 2012. The changing face of London. *The Economist*, Jan 28th, 2012.

Fotheringham, A.S., Brunsdon, C. & Charlton, M., 2002. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*, Chichester: John Wiley & Sons.

Goodhart, D., 2014. *Mapping Integration*, London: Demos.

Goodhart, D., 2013. White flight? Britain's new problem – segregation. *Prospect*, pp.30–31.

Gorard, S., 2009. Does the index of segregation matter? The composition of secondary schools in England since 1996. *British Educational Research Journal*, 35(4), pp. 639–652.

Gorard, S., 2011. Measuring segregation—beware of the cautionary tale by Johnston and Jones. *Environment and Planning A*, 43(1), pp.3–7.

Gorard, S., Taylor, C. & Fitz, J., 2003. *Schools, Markets and Choice Policies*, London: RoutledgeFalmer.

Greater London Authority, 2013. *Diversity in London*, London. Available at: http://data.london.gov.uk/datastorefiles/documents/2011-census-diversity-in-london.pdf.

Harris, R., 2014a. Evidence and trends: are we becoming more integrated, more segregated or both? In D. Goodhart (ed.) *Mapping Integration*. London: Demos. pp.15–23.

Harris, R., 2015. Measuring Changing Ethnic Seperations in England: a spatial discontinuity approach. *Environment and Planning A*, 46(9), pp.2243–2261.

Harris, R., 2011. Measuring segregation? A geographical tale. *Environment and Planning A*, 43(8), pp.1747–1753.

Harris, R., 2014b. "Sleepwalking towards Johannesburg"? Local measures of ethnic segregation between London's secondary schools, 2003 – 2008/9. In C.D. Lloyd, I.G. Shuttleworth & D.W.S. Wong (eds.) *Social-spatial segregation: Concepts, processes and outcome*. Bristol: Policy Press, pp.221–245.

Harris, R., Johnston R., Jones, K. & Owen, D., 2013. Are indices still useful for measuring socioeconomic segregation in UK schools? A response to Watts. *Environment and Planning A*, 45(10), pp.2281–2289.

Harris, R., Singleton, A., Grosse, D., Brunsdon, C. & Longley, P., 2010. Grid-enabling Geographically Weighted Regression: A Case Study of Participation in Higher Education in England. *Transactions in GIS*, 14(1), pp.43–61.

Hellen, N., 2013. Britons "self-segregate" as white flight soars. *The Sunday Times*, p.15.

Hong, S.-Y. & O'Sullivan, D., 2014. *Package "seg": A set of tools for measuring spatial segregation*, Available at: http://cran.r-project.org/web/packages/seg/seg.pdf.

Jakubs, J.F., 1981. A Distance Based Segregation Index. *Journal of Socio-Economic Planning Sciences*, 15, pp.129 – 141.

Jenkins, S., 2013. Sorry, Archbishop — but London is where the action is. *London Evening Standard*.

Johnston, R. et al., 2006. School and Residential Ethnic Segregation: An Analysis of Variations across England's Local Education Authorities. *Regional Studies*, 40(9), pp.973–990.

Johnston, R. et al., 2007. "Sleep-walking towards segregation?" The changing ethnic composition of English schools, 1997-2003: an entry cohort analysis. *Transactions of the Institute of British Geographers*, 33(1), pp.73–90.

Johnston, R., Forrest, J. & Poulsen, M.J., 2002. Are there Ethnic Enclaves/Ghettos in English Cities? *Urban Studies*, 39(4), pp.591–618.

Johnston, R. & Jones, K., 2011. A brief response to Gorard. *Environment and Planning A*, 43(1), pp.8–9.

Johnston, R. & Jones, K., 2010. Measuring segregation - a cautionary tale. *Environment and Planning A*, 42, pp.1264–1270.

Johnston, R., Poulsen, M. & Forrest, J., 2014. Increasing Diversity Within Increasing Diversity: the Changing Ethnic Composition of London's Neighbourhoods, 2001-2011. *Population, Space and Place*, in press.

Johnston, R., Voas, D. & Poulsen, M., 2003. Measuring spatial concentration: the use of threshold profiles. *Environment and Planning B*, 30(1), pp.3–14.

Katwala, S., 2013. The truth about London's "white flight." *New Statesman politics blog*.

Lee, B.A. et al., 2008. Beyond the Census Tract: Patterns and Determinants of Racial Segregation at Multiple Geographic Scales. *American Sociological Review*, 73(5), pp.766–791.

Lloyd, C.D., 2012. Analysing the spatial scale of population concentrations by religion in Northern Ireland using global and local variograms. *Journal of Geographical Information Science*, 26(1), pp.57–73.

Massey, D.S. & Denton, N.A., 1988. The dimensions of residential segregation. *Social Forces*, 67(2), pp.281–315.

Merry, M.S., 2011. Does Segregation Matter? In Bakker, J., Denessen, E., Peters, D. & Walraven, G. (eds.) *International perspectives on countering school segregation*. Apeldoorn, NL: Garant, pp. 249–260.

Merry, M.S., 2013. *Equality, Citizenship, and Segregation: A Defense of Separation*, London: Palgrave Macmillan.

Mitchell, J.E., 2002. Branch-and-Cut Algorithms for Combinatorial Optimization Problems. In Mauricio G. C. Resende, M.G.C. (eds.) *Handbook of Applied Optimization*. Oxford: Oxford University Press, pp. 65-77.

Mitchell, R. & Lee, D., 2014. Is there really a "wrong side of the tracks" in urban areas and does it matter for spatial analysis? *Annals of the Association of American Geographers*.

Monge, G., 1781. *Mémoire sur la théorie des déblais et de remblais. Histoire de l'Académie Royale des Sciences de Paris, avec les Mémoires de Mathématique et de Physique pour la même année*, pp. 666-704.

Morgan, B.S., 1983. An Alternative Approach to the Development of a Distance-based Measure of Racial Segregation. *American Journal of Sociology*, 88(6), pp.1237 – 1249.

Morrill, R.L., 1991. On the Measure of Geographic Segregation. *Geography Research Forum*, 11, pp.25–36.

Nightingale, C.H., 2012. *Segregation: A Global History of Divided Cities*, Chicago: University of Chicago Press.

O'Brien, N., 2012. Another country. *The Spectator*. Available at: http://www.spectator.co.uk/features/7779258/another-country/ [Accessed July 22, 2014].

Peach, C., 2009. Slippery Segregation: Discovering or Manufacturing Ghettos? *Journal of Ethnic and Migration Studies*, 35(9), pp.1381–1395.

Poulsen, M., Johnston, R. & Forrest, J., 2001. Intraurban ethnic enclaves: introducing a knowledge-based classification method. *Environment and Planning A*, 33(11), pp.2071–2082.

Reardon, S.F. & O'Sullivan, D., 2004. Measures of Spatial Segregation. *Sociological Methodology*, 34, pp.121–162.

Sassen, S., 2001. *The Global City: New York, London, Tokyo* 2nd ed., Princetown: Princeton University Press.

Shevky, E. & Williams, M., 1949. *The Social Areas of Los Angeles: Analysis and Typology*, Los Angeles: University of California Press.

Simpson, L., Gavalas, V. & Finney, N., 2008. Population Dynamics in Ethnically Diverse Towns: The Long-term Implications of Immigration. *Urban Studies*, 45(1), pp.163 – 183.

Waldorf, B.S., 1993. Segregation in Urban Space: A New Measurement Approach. *Urban Studies*, 30(7), pp.1151 – 1164.

Watts, M., 2013. Socioeconomic segregation in UK (secondary) schools: are index measures still useful? *Environment and Planning A*, 45, pp.1528 – 1535.

White, M.J., 1983. The Measurement of Spatial Segregation. *The American Journal of Sociology*, 88, pp.1008 – 1018.

Wong, D.W.S., 1993. Spatial Indices of Segregation. *Urban Studies*, 30(3), pp.559–572.

Wong, D.W.S., 1997. Spatial dependency of segregation indices. *The Canadian Geographer*, 41(2), pp. 128–36.

Wong, D.W.S., 2003. Implementing spatial segregation measures in GIS. *Computers, Environment and Urban Systems*, 27, pp.53–70.